CENTRE FOR HISTORY AND ECONOMICS

The Digitization of History
Notes from Google Books discussion, 19 July 2007

Attending: Leigh Denault (Cambridge/CHE), Anthony Grafton (Princeton), Jason Hanley (Google), Robert Watson (Cambridge/CL); meeting took place at Google's London office.

In July 2007, Robert Watson and Leigh Denault (Centre for History and Economics, Cambridge), and Professor Anthony Grafton (Princeton) met with a member of the Google Book Search team at Google's London offices to discuss the implications of large-scale digitization projects on the practice of history. The wide-ranging conversation covered issues from the technology of scanning and the management of large digital archives to the social implications of information accessibility. We have published the conversation here as an introduction to some of the issues facing commercial and academic digitization projects today.

Jason Hanley's background:

JH works with publishers directly, both with the existing partnered publishers (currently in excess of 10,000 total) and with publishing houses considering joining the project. STM (Scientific, Technical and Medical) publishing is his specialty. He formerly worked with Elsevier, on the *Lancet* and other STM publishing projects.

History & Philosophy of Google Books:

JH related the possibly apocryphal tale that the Books project actually predated the search engine, with Google beginning as a project to produce an electronic index for Stanford University library—in order to find a test data set, they used the world wide web, which at that time could be stored on a small number of computers.

Google famously came up with an idea or philosophy first ('organising information') and a business plan second, and this remains true of their current projects (Google Maps, You Tube etc.). Advertising has turned out to be the basis of their revenue, but this was not an expected outcome: Google tends to develop the technical ideas before thinking about marketing or social impact, a fact which has serious implications for how the Books project will affect scholarship.

JH made the distinction between 'accessible' content vs. 'free' content. Google's mission statement is to organise information, and they are therefore interested primarily in accessibility and 'discoverability' rather than in 'freeing' copyright. He observed that most information is not yet on the Internet. They see their model in this as the academic journal, which provides access to some content under copyright (article abstracts, tables of contents, author biographies etc.) while restricting access to full content to subscribers. Google however does not plan to offer full content: they see themselves as indexers and 'information organisers' rather than content providers.

Google is trying to provide a unified search interface across substantially different information sources: books, mapping, Google Scholar, video, etc. JH stated that Google was not 'in competition with Amazon,' and that they instead targeted an audience searching for information rather than an end product.

Content for the Google Books project is sourced, according to JH, in two ways:

1) From a publisher, under licensed agreement, an arrangement which usually includes provision for a limited sampling of the content online, capped overall but driven by the search query. This feature distinguishes Google Books search capabilities from the Amazon 'look inside' model, in that the book content returned is tailored to the user's search data, rather than a fixed excerpt, and thus more akin to actually browsing through a book in a bookstore.
2) Through the library programme, which started with five participating libraries and has now reached 26.

Here JH drew a graph on the whiteboard, courtesy of Tim O'Reilly: http://radar.oreilly.com/archives/2005/11/oops_only_4_of_titles_are_bein.html, demonstrating that of 32 million books published about 5% are currently in copyright and 20% clearly in the public domain (PD). This leaves 75% of these 32 million books whose copyright status is in flux, commonly called 'orphan books' because their copyright owners are difficult to track down (and, having in many cases inherited the copyright, may be unaware that they are indeed the copyright holders). This disputed 75% is what Google Books is most interested in putting online in searchable form. Their hope is that the 'snippet' viewer, displaying non-contiguous stripes of text in response to specific search terms, falls under the 'fair use' clauses of copyright law. We discussed copyright as a 'temporal beast'—not something fixed forever. Of particular concern was that the legal system has not yet 'caught up' with the digital age, leading to scenarios where libraries are legally unable to preserve some of their collection by making digital copies (especially audio media). The legal rule of thumb generally being that anything which causes economic damage to the copyright holder goes against 'fair use,' JH suggested that as Google Books in fact promoted sales of these 'orphan books', they were not transgressing copyright.

JH further suggested the digitisation debates are being replayed as different disciplines and genres are moving online. The STM market has already migrated to largely digital content, a move which, while it sparked controversy (the Elsevier bundling debates) is now largely perceived as having been beneficial to the disciplines and to accessibility.

JH noted that Google has been called 'arrogant' and 'audacious' for the comprehensiveness of their ambition, but felt that many of their critics had not actually engaged with the project's actual goals. In some theoretical sense the goal is to organize 'all information'—though given that some books don't survive and others are fragile, this will never be actualized.

AG here suggested that Google Books may indeed have outgrown the original intent of its designers, in that it was now a first port of call for scholars and students performing bibliographic searches, and perhaps the only means of accessing some information about recently published academic books and articles in developing countries. JH responded that Google was interested to hear about this kind of use, since they had never imagined that Google Books would become a scholarly tool, but saw it more as a general interest indexing service. AG suggested that the academic world was hardly a niche market re: Google's stated goals, particularly since with the phenomenal rise in access to University-level education, academics were educating the next generation of Google's book browsers and potential book buyers.

LD then asked about language and access, pointing out that Google Books might be the only means of accessing secondary content by scholars and students in developing

countries, and asking whether Google's comprehensive ambitions included expanding into non-Western languages or indeed significantly beyond the Anglophone world. JH asked whether LD was implying that Google had a social responsibility, reiterating his point that the project was never intended to replace library or trade bookstore access to material, but simply to index such materials. (NB we returned to this theme later in the discussion of redundancy/preservation.) Regarding languages, JH noted that the French had launched an early protest of the Google Books project as an exclusively Anglophone endeavour, but that the truth was that as an Anglo-American-based company, they had simply started with what they knew. As the project expanded and they added partners in different regions (he said that there was already significant participation in Europe, East Asia, India and the Middle East), Google hoped to be able to add more language support. Google currently indexes material in seven European languages, and in Japanese and Chinese. Google depends heavily on its publishing and library partners to provide expertise for non-Roman fonts, especially for script languages such as Arabic language, which have seen significantly less focus in past OCR research and products. JH also noted that the system used for English OCR has been put in the public domain.

JH however suggested that the degree to which a country engaged with the project was dependent not just on technical or linguistic issues, but also on other internal social and economic factors. Denmark, for example, with a very small expatriate community, a well-organised and thriving bookshop system whose proprietors are considered authoritative sources of information, has not jumped on the digitisation bandwagon. Germany provides another example of the country which seems to have all of the right social ingredients to embrace online book shopping/searching, but due to cash-centred domestic shopping habits and low use of online shopping (considered a gateway metric predicting use of digital resources more generally) German companies and individuals are also less likely to use Google's projects.

Thus there are social as well as technical obstacles to expanding the project beyond the Anglophone world and beyond its current mission of 'discoverability' over content provision.

Partners, Content and 'Containers'

RW asked if JH could then explain how Google works with particular partners to put content online. JH explained that he usually works with publishers, many of whom have the past decade or so of their assets in digital form, be it in PDF or Quark. Finding archival material in digital form is more rare, and OCR and scanning is required to make these books available online.

RW asked about whether Google provided any guarantees about the durability of the archive they were creating for their partners. JH said that most appeared to rely on Google's status as a company with likely staying-power rather than on explicit guarantees about the material, and typically had their own archives for published materials. Any partner, both library and publishing, is allowed to retain their own copies of the scans which Google creates for internal use, but many of their partners do not have the facilities to store the scans and often are content to leave them with Google. Google, JH noted, does not see itself as archivist but recognizes the clear alignment of goals. JH said that Google had been working closely with the Bodleian on particular library collections with an eye to preservation, but that most of their partners saw the Google Books project as provide an alternative kind of marketing rather than a durable archive or backup of their assets. Their reliance on redundancy to minimise latency for

search users has by default meant that Google has a lot of de facto or occult backups of their material, but they have not specifically instituted a policy of 'preservation' of their content. Google recognizes—but of course does not take responsibility for—the fact that libraries will need to develop very substantial technical infrastructure in the near and longer-term future.

This led to JH's point that online content is more generally seen as transient, a fact demonstrated by the difficulty of citing such material. He wondered whether copyright libraries might have a responsibility to archive the web, as well as whether Google should reassess their role in preservation. The web has not up to now been viewed as a scholarly tool by Google, but rather as a generally democratising trend in information accessibility.

LD asked here about the possibility that Google Books might ultimately expand into being a provider of content, as well as bibliographic information, and whether they might then consider, along the democratising theme, providing tiered access to different countries and academic institutions. JH said that right now, such expansion or access schemes would be entirely dependent on the wishes of their partner publishers and libraries. Google, he stressed, was still ultimately bound by publishers' models. He speculated that if, three years on, publishers found that they were deriving revenue from digitised content, then something more ambitious might be possible, but that copyright of those digitised images would be key and would probably still bind the digital world to the print models. Google is however focusing on building and expanding their relationships with libraries and publishers in the hope that more and more content can be put online and into broader use.

Google however has no intention of replacing libraries or bookstores, and was, JH stressed again, never built to be a scholarly tool. JH reiterated that as a technology, a book has a lot going for it, and has in fact remained an unmodified information 'container' for several hundred years. AG said that the while the fit between scholarly and public needs might not be exact, Google was still being used heavily by a scholarly audience.

JH said that their US library partners had a disclosure agreement specifying how all digital content would be used and stored, and that the Bodleian material now hosted by Google has a similar agreement. The libraries are allowed to retain copies for their own use. Google provides a similar scheme to publishers, allowing them to control how much information or content is available via a click-through interface, and whether or not to include advertisements on the page. JH referenced Nicholson Baker's *Double Fold,* suggesting that the jump to digital was in many ways detrimental to the preservation needs of print, and that digitisation should never be seen as a replacement for print materials. AG brought up the example of the microfilming of four-colour lithographed newspapers in black and white, after which the originals were scrapped, ensuring that future readers would no longer be able to see these texts as they had been printed

RW and LD asked about accessibility from the perspective of disabled users as well as for users with low-bandwidth requirements, and JH said that in fact both of these had been a major focus of Google. Google has resisted the 'Web 2.0' model (which brought in heavy use of images and high-latency programming) both for technical and social reasons. Google has made the text layer (in addition to the scanned image) of all public domain books available for disabled access to their content (as text-to-speech readers are able to interpret the plain text layer but not the graphical image content).

AG and RW then asked a series of questions dealing with issues surrounding the scanning of material from libraries. JH stated that Google Books currently has one million books in their index for searching, and 10,000 publishing partners globally. On the question of how Google Books deals with multiple editions of a text, JH was not certain but thought that there would probably be little effort made at this stage to distinguish between different editions or indeed even to prevent the scanning of a single edition more than once at multiple partner sites. AG talked about the value of textual comparison in studies of the history of the book, or indeed for cultural and intellectual historians of all stripes; JH responded by explaining that Google saw this as a long process, in which they were making up standards and practices as they went and in which there would be a great deal of trial and error. Their main priority was simply to get as much material online in 'accessible' form as possible, and they understood that errors and problems with the texts would need to be addressed in the future.

How do you measure success? Google Books and their Critics

JH stressed several times that Google was more of a concept than a business model, and that their current 'gold mine,' i.e. advertising in online, print and radio media, was very much a second-generation development for the company. Google tends to work to a 'cool app' model, in which a fun or useful idea comes first and the enterprise model comes decidedly later.

In regard to Google's vocal critics, JH said that it was not as if Google has 'form for devious behaviour.' Rather, he said, they do have 'form for working with projects that do not make money,' simply because the company believes they are important. When AG and RW asked by what metrics Google might measure success, JH said that it was definitely not by the classic P & L business model, but rather through the comprehensiveness of Google indexing, through Google's expansion into new areas, and through perfecting search algorithms to return information 'containers' ranked well by relevance to the user's query. RW inquired about whether or not Google kept and used existing meta-data about scanned material, including catalogue information, physical location, etc. For the Books project, JH noted that the next big change would be a separate 'more about this book' page for each indexed item, ultimately to contain abstracts, images, relevant maps and statistical data etc. They are also interested in producing a 'browsing' function which would turn up related books. Amazon, with its 'statistically improbable phrase' feature, may be ahead here. (AG and RW here noted that shelf-browsing functionality is crucial to student and scholarly use of Google Books, to retain the serendipity factor which drives so many academic searches in open-stack university libraries).

In response to our queries about future relationships between Google, academics and professional archivists and librarians, JH said that Google already was able to make available specialised tools in academic departments and libraries to customise searches, to teach students to use the new digital resources properly. AG said that it was crucial to keep in mind how a lack of dialogue between scholars and computer scientists had resulted in some fairly fundamental public misunderstandings of the Google project, and that this is especially sensitive given our previous discussion of preservation issues and whether librarians should see Google Books as a viable long-term solution to their maintenance and budgetary problems (a role Google does not see itself fulfilling!).

Books from libraries vs. books from publishing partners are scanned using a 'totally different' process, and it is in part this difference which leads to lower quality in library

scans. The hope is however that Google can ‘launch early’ and make iterative changes, re-running OCR on previously scanned images and re-scanning problematic scans as technology improves.

Google usually receives more criticism than positive feedback, JH noted. While they did perform extensive usability and Human-Computer Interaction (HCI) studies, which led directly to the three major iterations of Google’s layout and user interface, Google would be interested in continuing the dialogue with Cambridge, and he suggested a contact with the library programme who could answer questions about library partners in more depth.

L. Denault and R. Watson, 17 October 2007